Baba Ghulam Shah Badshah University Rajouri (J&K)

(SYNOPSIS FOR REGISTRATION FOR DEGREE OF DOCTOR OF PHILOSOPHY IN APPLIED MATHEMATICS)

Proposed Topic	:	"**A Study of Climatic and Economic Data Mining Using Wavelet Methods**"
Name of the Candidate	:	Mr. Mudassar Rashid Lone
Name of the Supervisor	:	**Supervisor:** Dr. Zaheer Abbas Department of Applied Mathematics, Baba Ghulam Shah Badshah University, Rajouri. **Co-Supervisor:** Prof. A.H.Siddiqi Department of Mathematics, Gautam Buddha University, Greater Noida,U.P.
School/Department	:	Applied Mathematics, School of Mathematical Sciences & Engineering.
Date of Submission	:	

_____
Signature of the Candidate

_____ _____

Signature of the Supervisor Signature of the Co-Supervisor

1. Background

Introduction:

The data are usually in the form of signals, images. These data sets arise in diverse fields such as financial markets, meteorology, medical imaging, physics, chemistry, material sciences, astronomy and bioinformatics. They can be obtained from simulations, experiments or observations. Data mining is the process concerned with uncovering patterns, associations, anomalies, significant features and structures in data. It is a multidisciplinary field borrowing and enhancing ideas from different domains including image processing/signal processing, machine learning, optimization, high performance computing information retrieval and computer vision. Extraction of novel, useful and understandable patterns from large collection of data is known as **data mining or knowledge discovery** in data base.

The important ingredients of data mining are

1. Characterization of different classes of data
2. Management of data
3. Dimensionality reduction
4. Denoising
5. Data compression
6. Fusion of data
7. Enhancement of data
8. Pattern recognition in the data
9. Extracting features in the data

10. Finding missing elements in the data

11. Prediction

12. Visualization

The wavelet methods have been applied to study some of the above mentioned features of data mining. The theory and applications of wavelets have been vigorously studied by physicists, mathematicians and engineers alike in the last decade. The name wavelet means small waves and in brief, a wavelet is an oscillation that decays quickly. Formally, we define wavelet as :

If $\psi \in L^2(\mathbb{R})$ satisfies the admissibility condition $C_\psi = \int_{-\infty}^{\infty} \frac{|\widehat{\psi}|^2}{\omega} d\omega < \infty$, then ψ is called the basic wavelet and relative to every basic wavelet ψ, the wavelet transform of a function f on $L^2(\mathbb{R})$ is given by

$$\left(W_\psi f\right)(b,a) = |a|^{\frac{-1}{2}} \int_{-\infty}^{\infty} f(t)\overline{\psi\left(\frac{t-b}{a}\right)} dt$$

where $a, b \in \mathbb{R}, a \neq 0 \; and \; f \in L^2(\mathbb{R})$.

A wavelet transform decomposes a signal into several groups (vectors) of coefficients. Different coefficient vectors contain information about characteristics of the sequence at different scales. It may be observed that the wavelet transform is a prism which exhibits properties of signal such as points of abrupt changes, seasonality or periodicity.

Some of the important examples of wavelets are

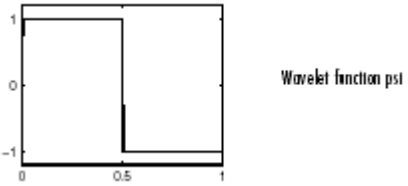
1. Haar wavelet:

The function defined below is called the Haar function.

$$\psi(x) = \begin{cases} 1 \quad , & 0 \leq x \leq \frac{1}{2} \\ -1 \; , & \frac{1}{2} \leq x \leq 1 \\ 0 \quad , & \text{otherwise} \end{cases}$$

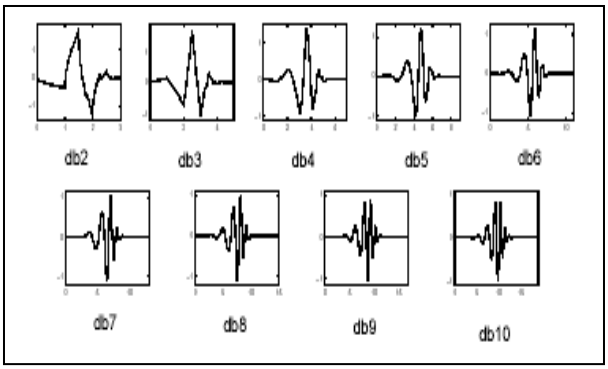
This is first and simplest wavelet and any discussion of wavelets begin with Haar wavelet. Haar wavelet is discontinuous, and resembles a step function.

Graphically:



2) Daubechies Wavelet

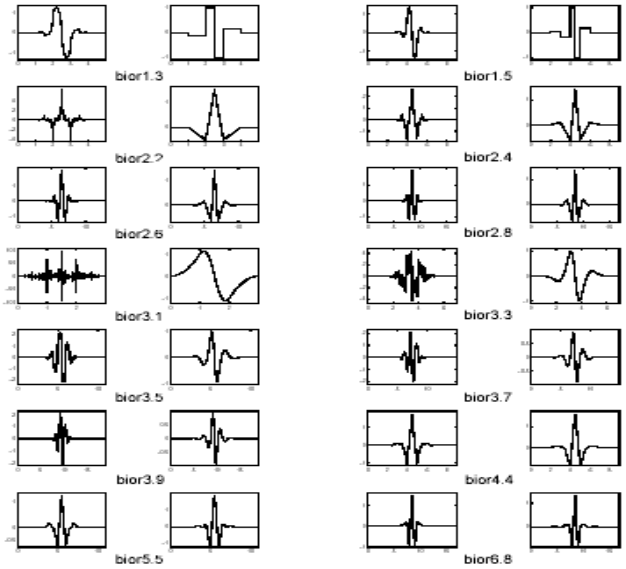
Ingrid Daubechies, one of the brightest stars in the world of wavelet research, invented what are called compactly supported orthonormal wavelets — thus making discrete wavelet analysis practicable. The names of the Daubechies family wavelets are written dbN, where N is the order, and db the "surname" of the wavelet. The db1 wavelet, as mentioned above, is the same as Haar wavelet. Here are the wavelet functions psi of the next nine members of the family:



3) Biorthogonal

This family of wavelets exhibits the property of linear phase, which is needed for signal and image reconstruction. By using two wavelets, one for decomposition (on the

left side) and the other for reconstruction (on the right side) instead of the same single one, interesting properties are derived.

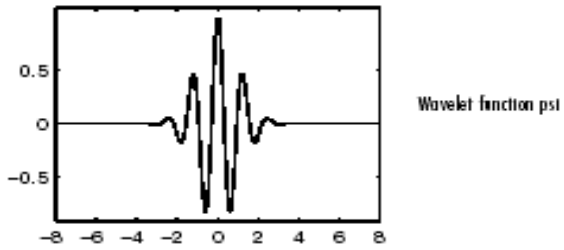


4) Morlet wavelet

A function defined as

$$\psi(x) = (e^{i\alpha t} - e^{\alpha^2/2})e^{x^2/2}$$

is called morlet or Gabor function.This wavelet has no scaling function, but is explicit.

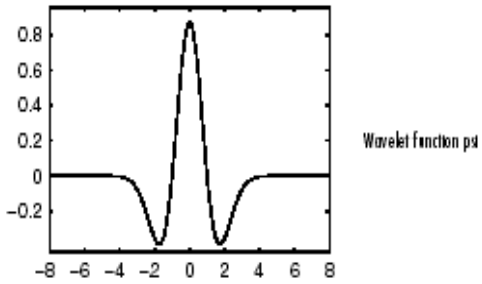


5) Mexican Hat wavelet

The function ψ defined as

$$\psi(x) = (1 - x^2)e^{-x^2/2}$$

is known as the Mexican hat function. Mexican Hat wavelet has no scaling function and is derived from a function that is proportional to the second derivative function of the Gaussian probability density function.

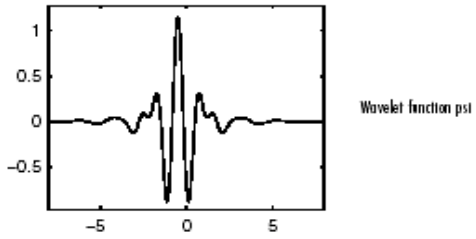


6) Meyer wavelet

The Lemarie-Meyer family of wavelets and scaling function are defined in the frequency domain and the family of wavelets is of the form

$\hat{\psi} = b(\xi)e^{i\xi/2}$, where, b is an even function on R.

Graphically:



In images, high amplitude wavelet coefficients indicate the position of edges, which are sharp variations of the image intensity and different scales provide the contours of image structures of varying sizes. Such multiscale edge detection is particularly effective for pattern recognition in computer vision.

There are several fields in which wavelets find its application. The major contributions are to the fields of:

(a) Medical sciences

(b) Bioinformatics and Computational Biology

(c) Environmental studies

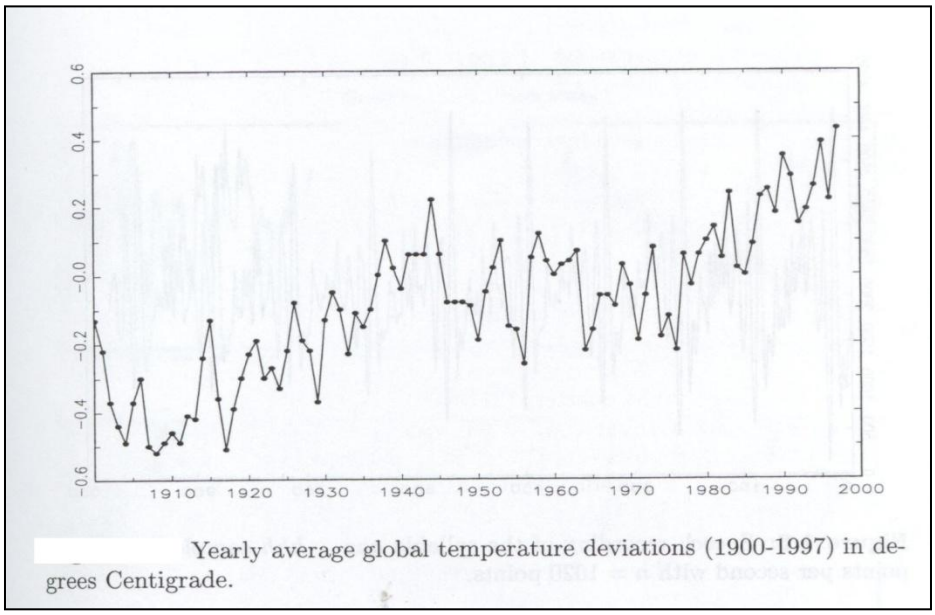
(d) Finance etc.

A wavelet transform leads to an additive decomposition of a signal into a series of different components describing smooth and rough features of the signal.For analyzing real-world problems time series analysis approach is adapted. A time series is a sequence of event values which occur during a period of time and the analysis of experimental data that have been observed at different points of time is known as time series analysis. In time series analysis approach we observe following properties of the time series (signal)

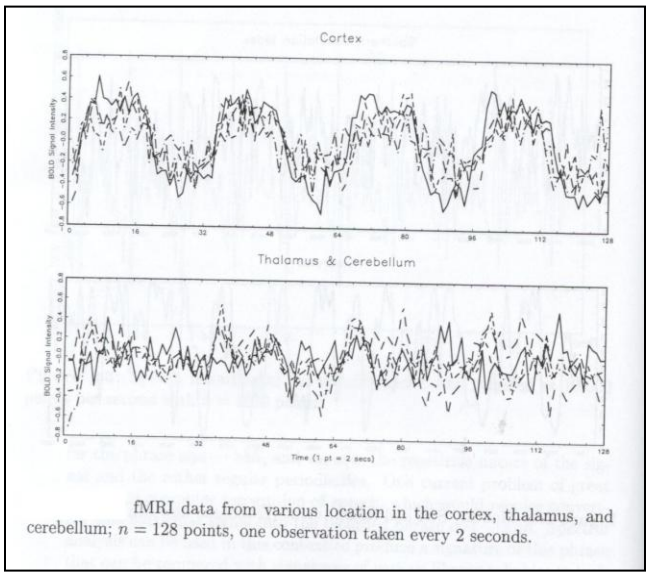
- Trends
- Business Cycles
- De-noise
- Self Similarity
- Compression
- Periodicities
- Abrupt changes
- Drift
- Forecasting / Prediction etc.

There are various situations where time series can occur such as:

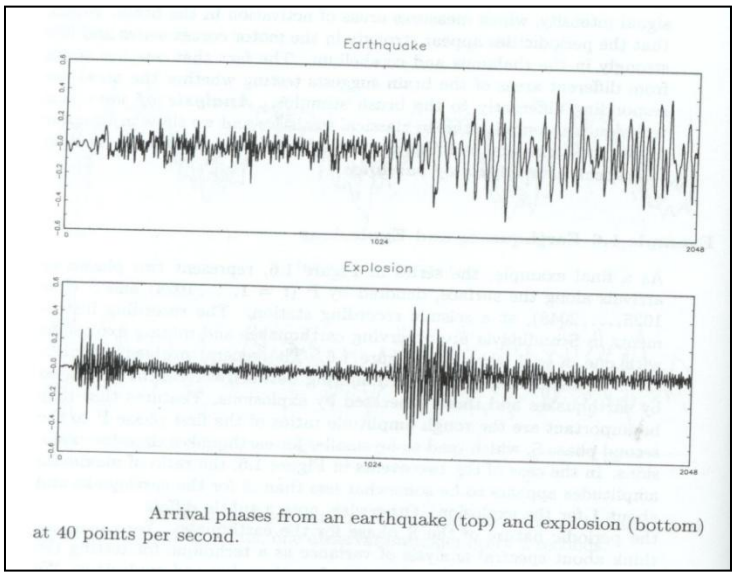
- In the field of economics/ banking system, where, one is exposed to stock market quotations or monthly unemployment figures or foreign exchange rates.

- In social sciences, where, population changing time series, such as, birth rate time series or university / school enrollments.
- In medical sciences, where, we need to study time series of influenza cases during certain period of time, blood pressure measurements traced over time for evaluating drugs used in treating hyper-tension.
- In Electrocardiogram (ECG) data and functional magnetic resonance imaging of brain, where, wave time series patterns study how the brain reacts to certain stimuli under various experimental conditions etc.
- 

Yearly average global temperature deviations (1900-1997) in degrees Centigrade.



fMRI data from various location in the cortex, thalamus, and cerebellum; $n = 128$ points, one observation taken every 2 seconds.

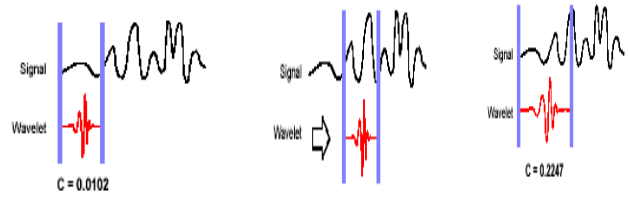


Arrival phases from an earthquake (top) and explosion (bottom) at 40 points per second.

Further, in engineering and environmental sciences we come across a lot of time series for

- Measurements acquired with change in the atmospheric layers
- Rain fall effecting the agricultural products
- Surface evaluation corresponding to wind generated waves measured near the shore areas of sea/lake/river
- Temperature variation and wind pressure
- Nuclear Reactor
- Ultra sound and vibrations
- Blood flow sounds, heart sounds and lung sounds
- Global warming
- Speech data
- Earth quakes
- Explosions etc

Wavelet Analysis of a time series is the study of previously mentioned properties of time series (signal), by breaking up of the signal (time series) into shifted and scaled version of the original (mother) wavelet. As mentioned earlier scaling a wavelet simply means stretching (or compressing) it. One can make a plot on which the x-axis represent position along the signal (time) the y-axis represents scale and the color at each x, y point represent the magnitude of each wavelet coefficients. See figure below



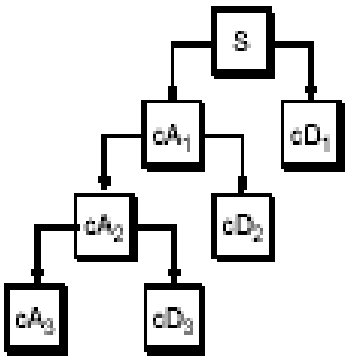
For many signals, the low frequency content is the more important part. It gives signal its identity. The high frequency content, on the other hand, imparts flavor or nuance. For example in human voice if the high frequency components are removed the voice sounds

different, but still tells what is being said. However if we removed enough of the low frequency components we hear gibberish.

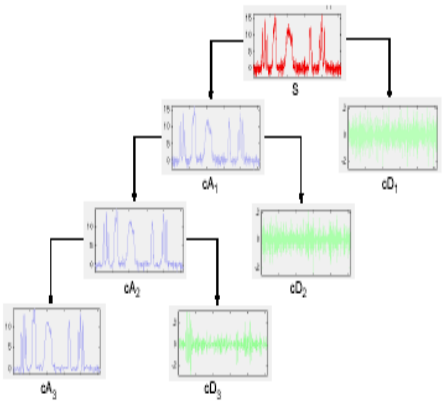
Thus in wavelet analysis we can divide signal in parts, one type called approximations and other called details.

The approximations are the high scale low frequency components of the signal, the details are the low scale high frequency components of the signal.

The decomposition process can be iterated, with successive approximations being decomposed in turn, so that one signal is broken down into many lower resolution components. This is called the wavelet decomposition tree.



Looking at a signal's wavelet decomposition tree can yield valuable information.



Since the analysis process is iterative, in theory it can be continued indefinitely. In reality, the decomposition can proceed only until the individual details consist of a single sample or pixel. In practice, we will select a suitable number of levels based on the nature of the signal, or on a suitable criterion such as entropy .

Literature review

T.Li, Q.Li, S.Zhu and M.Ogihara[16] have written a beautiful review in Application of Wavelet to Data Mining in 2003. There are 177 references in this review giving updated results. Recently there has been significant development in the use of wavelet methods in various data mining processes. However, there has been written no comprehensive survey available on the topic. The goal of T.Li et al. is to fill the void. First, T.Li et al.present a high-level data-mining framework that reduces the overall process into smaller components. Then applications of wavelets for each component are reviewd. Final conclusion is made by discussing the impact of wavelets on data mining research and outlining potential future research directions and applications. In 2009 C.Kamath[13], has written a very interesting book on scientific data mining published by a reputed publisher, SIAM, Philadelphia. This book contains contains 654 references including some on wavelet to data mining. In January 2011 Chaovalit, A Gangopadhyay, G. Karabatis and Z.Chen[10] have published a review of applications of discrete wavelet transform particularly to one dimensional signal. This review article contains 132 references. The article introduces a wavelet-based time series data analysis and provides a systematic survey of various analysis techniques that use discrete wavelet transformation (DWT) in time series data mining, and outlines the benefits of this approach demonstrated by previous studies performed on diverse application domains, including image classification, multimedia retrieval, and computer network anomaly detection. The topics discussed are time series data characteristics, discrete wavelet transform for dimensionality reduction, application of wavelet transform to noise removal, singularity deduction, similarity search, classification, Anomaly deduction, prediction have been studied and updated results are presented. A.H.Siddiqi et al. [24] have studied applications of wavelets to data related to oil industry, meteorological data of Saudi Arabia applying methods of wavelets and fractals. In this paper, wavelet concepts are used to study a correlation between pairs of time series of meteorological parameters such as pressure, temperature, rainfall, relative humidity and wind speed. The study utilized the daily average values of meteorological parameters of nine meteorological stations of Saudi Arabia located at different strategic locations. Besides obtaining wavelet spectra, A.H.Siddiq et al. also compute the wavelet correlation coefficients

between two same parameters from two different locations and show that strong correlation or strong anti-correlation depends on scale. The cross-correlation coefficients of meteorological parameters between two stations are also calculated using statistical function. For coastal to costal pair of stations, pressure time series are found to be strongly correlated. In general, the temperature data are found to be strongly correlated for all pairs of stations and the rainfall data the least. Important references are also given in the talk of P.Manchanda [19].

Study area

Applications of Wavelet Methods for

- Detecting discontinuity and breakdown points
- Detecting long term evolution
- Detecting self similarity
- Identifying pure frequencies
- Suppressing signals
- De-Nosing signals
- Compressing signals

2. Aims & Objectives

The researcher intends to undertake the work with the following aims and objectives:-

1. To study Dimension Reduction and Denoising using wavelet method for metrological, pollution and economic data of Jammu and Kashmir.
2. To study Similarity search and Prediction using wavelet method for metrological, pollution and economic data of Jammu and Kashmir.
3. To carry out case studies using metrological, economic, pollution data of district Rajouri in particular and Jammu & Kashmir in general .

3. Material proposed to be used for the investigation

There are two primary requirements for the proposed study, *datasets* and *simulation packages*. The datasets for the proposed work are mostly available online and are free for research purposes. The intended work will use the data available from the below mentioned sources.

1. Organizational/Institutional Websites.
2. Digital Citation Libraries.
3. Web Links.

For simulation purposes,packages like MATLAB will be used and, where necessary required software/simulation package may be downloaded from freely available sources from the Internet for the study at hand.

4. Methodology:

Figure 1 outlines the key processes in the extraction and mining of various data .

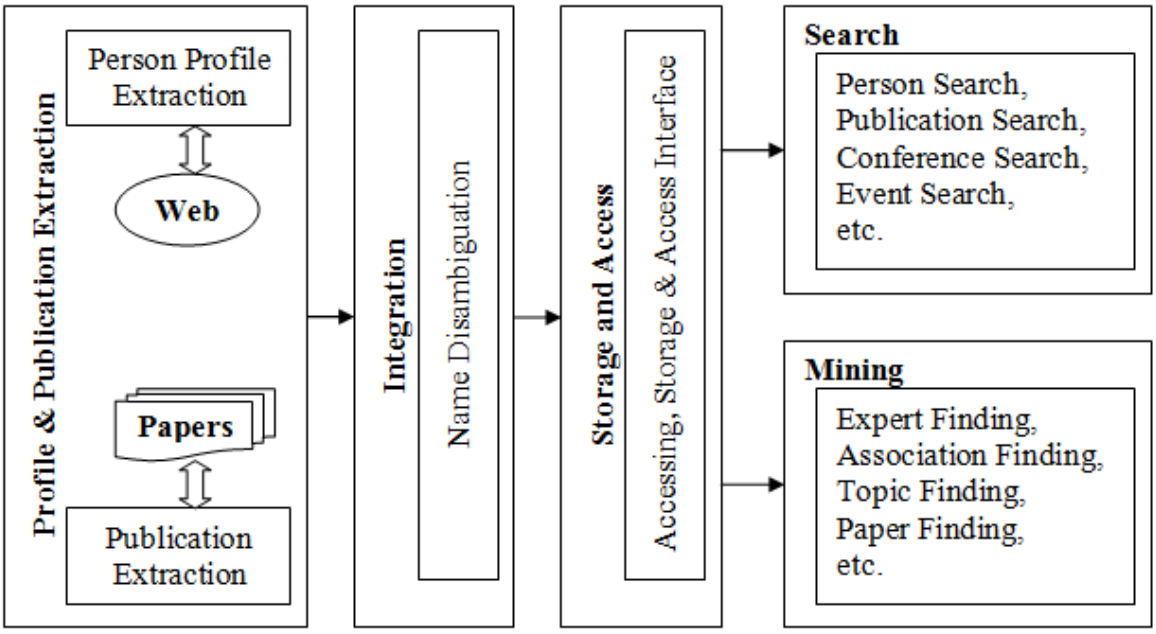


Figure 1: ***Steps in the Extraction of an Academic Social Network.***

Tools and Techniques proposed to be used for the proposed work:

Wavelet decomposition on the observed data facilitates quantitative or qualitative analysis of data, by describing features of the data, either through numerical or visual representation. Wavelet analysis software generally consists of either packages based on graphical user interfaces (GUIs), or packages built for scripting/programming languages. Packages like *MATLAB*, ANFIS etc. will be used for the purpose.

Main idea is to substrate the prediction task of the original time series of high variability by the prediction of its wavelet coefficients on different levels of lower variability, and then using Matlab code with ANFIS for final prediction at any instant of time. Since most of the wavelet coefficients are of lower variability we expect the increase of the total prediction accuracy. Daubechies db wavelet at any level may be used for decomposing the given time series signal. The basic idea to use the wavelet transforms and predict the data by neuro fuzzy for individual coefficients of wavelet transform will remain as same.

Traditional data mining techniques such as *clustering*, *classification*, *association rule mining*, and *visualization* will also be used. In data mining, classification and clustering can be used to create different classes of users, the difference between the two is that, in *classification* classes are predefined (*supervised*) and in clustering they are not predefined (*unsupervised*). *Association rule mining* technique can be used to discover direct or indirect relationships between data entities. *Visualization* is a special technique to present data and information in graphical, understandable manner and plays an important role in data mining.

Action Plan and Time Line:

Of the steps in extraction of a data discussed earlier, storage, access and searching have got quite a lot of attention from the research community over the years. Well reputed and efficient methods for the same exist. The emphasis of this proposed work shall be *extraction*, *integration and prediction*. It is proposed to carry out the research work over a period of three years and the tentative work plan for each year is as follows:

First Year:

During the first year the main task would be to study and understand various techniques proposed so far in the proposed field of investigation. What methodologies and techniques have been followed in their design and what are the loopholes and shortcomings in these approaches. In addition all efforts will be made to have a comprehensive literature collection on different aspects of data mining.

Second Year:

Once the study of various conventional techniques proposed so far for is over then, the few good evaluation parameters will be selected for comparative study of different methods. Case studies will be carried out using metrological, economic, pollution data of district Rajouri in particular and Jammu and Kashmir in general. The proposed system will be subjected to experimental testing for their performance. Based on the evaluation parameters decided, an effort will be made to propose the best possible system/solution.

Third Year:

The resultant system/solution will be tested for performance with the available benchmark datasets and comparison shall be drawn with the baseline methods.

Final Reporting of facts and details of background will be presented in the form of book as thesis for completion of research work. Around six months of the third year will be devoted for thesis writing and checkup.

5. Possible Outcome of the investigation

Important research papers related to prediction, denoising, dimension reduction, similarity search etc for real world data.

6. References:

[1] **Abramovich F. et al.** (2000), Wavelet Analysis and its Statistical Applications, The Statistician, Vol.1, pp. 1-29.

[2] **Addison S.** (2002)**,** The Illustrated Wavelet Transform Handbook: Introductory Theory and Application in Science, Engineering, Medicine and Finance, Institute of Physics Publication, Publishing Bristol and Philadelphia.

[3] **Agniar L. et al.** (2010), Oil and the Macro economy: Using wavelets to analyze old issues, Empire Ecom, Springer.

[4] **Akansn AN. et al.** (2010), Emerging Applications of Wavelets: A review, Physical communication, Vol.3, pp. 1-18.

[5] **Bonchonev I. et al.** (2002), Recovery of Volatility Coefficient by Linearlization, Quantitative Finance Vol.2, pp. 257-363.

[6] **Chen P.**, Special Volume on Data Mining, Springer Science Business, 2010, pp.174

[7] **Chiann C. et al.** (1999), A Wavelet Analysis for Time Series, Journal of Nonparametric Statistics, Vol. 10, pp. 1–46.

[8] **Duri AD. et al.** (2009), Analysis of Brain Electrical Topography by Spatio-Temporal Wavelet Decomposition, Mathematical and Computer Modelling Vol. 49, pp.2224-2235.

[9] **Furati KM. et al.** (2006), Mathematical Models and Methods for Real World Systems, Chapman & Hall /CRC, Taylor and Francis Group, Singapore.

[10] **Gangopadhyay A. et al.** (2011), Discreet wavelet transform based time series analysis and mining, ACM Computing Surveys, Vol. 6, pp.1-32

[11] **Han J. and Kamber M.** (2000), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.

[12] **Heinlein P.** (2003), Integrated Wavelets for Enhancement of Micro Calcifications in Digital Mammography, IEEE Transactions on Medical Imaging, Vol. 22, pp.402-413.

[13] **Kamath C.** (2009), Scientific Data Mining- A practical Prospective, SIAM, Philadelphia.

[14] **Laine AF.** (2000), Wavelets in Temporal and Spatial Processing of Biomedical Images, Annu Rev Biomed Eng, Vol.1, pp. 511-550.

[15] **Lia S. et al.** (2004), Data Mining to Aid Policy Making in Air Pollution Management, Expert Systems with Applications, Vol. 27, pp. 331–340.

[16] **Li T. et al.** (2000), A Survey On Wavelet Applications In Data Mining, SIGKDD Vol.42, pp. 49-67.

[17] **Manchanda P. et al.** (2007), Mathematical Methods for Modeling Fluctuations of Financial Time series, Journal of Franklin Institute, Vol.344, pp. 613-636.

[18] **Manchanda P. et al (**2002), Current Trends in Industrial and Applied Mathematics, Anamaya, New Delhi.

[19] **Manchanda P.** (2011), Prediction and Extraction of Events using wavelet Method, ICIAM 2011.

[20] **Megalooikonomou V. et al.** (2000), Data Mining in Brain Imaging, Statistical Methods in Medical Research, Vol.9, pp. 359–394.

[21] **Miller R.J et al.** (2002), Similarity Search over Time-Series Data using Wavelets, In ICDE 2002.

[22] **Osowsk S. et al.** (2007), Forecasting of the Daily Meteorological Pollution using Wavelets and Support Machine, Engineering Applications of Artificial intelligence, Vol. 20, pp.745-55.

[23] **Shumway R. et al.** (2000), Time Series Analysis and its Applications, Springer Vol. 22.

[24] **Siddiqi A.H, (**2012), Applications of Wavelet Methods to Oil industry and Meteorology, edited Vol. American Institute of Physics.

[25] **Walden A.T. et al**. (2000), Wavelet Methods for Time Series Analysis, Cambridge University Press.

[26] **Zhao J. et al.** (2003), Detecting Region Outliers in Meteorological Data, GIS'03, New Orleans, Louisiana, USA.